Jornada de Seguimiento de Proyectos en Tecnologías Informáticas, 2003
Programa Nacional de Tecnologías de la Información y las Comunicaciones

Technological platform for the semantic web: ontologies, natural language analysis, and e-commerce. TIC2001-2745

Asunción Gómez Pérez[1], Guadalupe Aguado de Cea[2],
Inmaculada Álvaro de Mon y Rego[3], Mariano Fernández López[4],
Óscar Corcho García[5], David Manzano Macho[6],
Atonio Pareja Lora[7], Ángel López Cima[8]
Laboratorio de Inteligencia Artificial

Abstract

The goal of this project is to provide core technology for the semantic web. Our aim is to develop an ontology-based platoformfor allowing users to query e-commerce applications by using natural language, performing the automatic information retrieval from web docuemtns annotated with ontological and linguistic information. In particular, our aims are:

1. To explore semi-automatic ontology construction by reusing ontologies already implemented in semantic web languages and e-commerce standards and joint initiatives.
2. To design and implement an environment for hybrid annotation (ontological and linguistical) of web documents
3. To build a new interface (henceforth "semantic portal"), which permits consulting the contents of web pages in any domain taking into account the information supplied by an existing ontology (previously built)
4. Ontology-based system for querying and information retrieving from annotated web documents in the cultural shows domain

Keywords:semantic way, ontologies, annotations

1 Aims of the project

[1] e-mail:asun@fi.upm.es
[2] e-mail: lupe@fi.upm.es
[3] e-mail: ialvarez@etsii.upm.es
[4] e-mail: mferandez.eps@ceu.es
[5] e-mail: ocorcho@fi.upm.es
[6] e-mail: dmanzano@delicias.dia.fi.upm.es
[7] e-mail: apareja@sip.ucm.es
[8] e-mail: alopez@delicias.dia.fi.upm.es

This project aims specifically to develop a platform based on ontologies for browsing and retrieving automatically textual information from web pages which are annotated with linguistic and ontological information; this textual information is to be used in a semantic web environment. We intend to add new functionalities/applications to the development platform of the WebODE ontologies and adapt it so as to be a useful platform within the context of the semantic web.

With this purpose in mind, in this project we have explored what is the contribution of ontological engineering to natural language processing and how natural language processing is needed to carry out some specific activities (autotomatic ontology construction) of the ontology development process.

In our proposal we have identified four sub-plans in order to achieve the aforementioned aims.

Subplan I To explore semi-automatic ontology construction

The main goal of this subplan is to develop technology that ease the construction of new ontologies by means of reusing existing ontologies and natural languages techniques. The goal is to develop technology for importing ontologies implemented in semantic web languages (RDF (S), DAML+OIL and OWL) into the WebODE platform and being able to extend such ontologies using natural language analysis techniques.

Subplan II OntoTAg: Hybrid Annotation Environment of web pages.

We pretend to build an annotation model for web pages where the morphological, syntactic and semantic issues are combined; this model should be compatible with the coding standards used in the Web. The annotation environment is independent of the ontology and is now being built as an additional service of WebODE.

Subplan III OntoConsult. Natural language interface based on ontologies.

Its goal is to build a new interface (henceforth "semantic portal"), which permits consulting the contents of web pages in any domain taking into account the information supplied by an existing ontology (previously built).

Subplan IV OntoAdvice. A system based on ontologies for browsing and collecting information from web pages annotated in the entertainment domain.

We will integrate the technology produced in the previous subplans in a prototype on the entertainment domain.

2 Level of success achieved in the project

In this section we summarize the main goals achieved, the main diffiulties found and the main results for the four subplans.

Subplan I **To explore semi-automatic ontology construction (ontology learning) by reusing ontologies already implemented in semantic web languages and e-commerce standards and joint initiatives.**

We aim to build domain-independent technology for the ontology building process.
For this reason, we have developed translators that import (into the WebODE platform) ontologies and e-commerce standards implemented in semantic web languages; we have also designed a tool, which is now been implemented, that permits building ontologies semi-automatically by using techniques of natural language. In the last two years, several activities within this sub-plan have been carried out and the results obtained are the following:

- State of the art on the e-commerce standards, including: UNSPSC, rosset-net, e-cl@s, etc.
- Development of translators that permit extracting semi-automatically contents from the information sources aforementioned. A new tool, called WebPicker, has been built specifically for the e-commerce domain.
- We have built translators that import ontologies implemented in semantic web languages (RDF(S), DAML+OIL, and OWL) into the WebODE platform.
- WebPicker and the translators have been integrated in the WebODE platform.
- We have developed a translator that permits exporting ontologies in semantic web languages: RDF(S) and OWL.
- We have also designed a module for learning ontologies; this module permits to extend existing domain ontologies with new information obtained by analysing domain texts with natural language techniques.
- The integration of the Spanish EuroWordNet into WebODE is now in progress. We pretend that EuroWordNet be used for ontology learning.
- We have already started to implement a prototype that permits extending the taxonomies of an ontology.

The following tasks are not finished yet:

- EuroWordNet integration.
- The prototype for extending the taxonomy semi-automatically and integrating it into WebODE.
- Technological Tranferer.

No deviations are expected in this sub-plan.

Subplan II **OntoTAg: Hybrid Annotation Environment of web pages.**
In this step, we aim to build a hybrid annotation model for web pages where the morphological, syntactic and semantic annotations are combined, according to liguistic annotation standards and recommendations and being compatible with the coding standards handled in the web. The annotation environment should be independent of the ontology and is now being built as a WebODE additional service.
In the last two years we have carried out the following activities with the following results:

- State of the art of models and platforms of linguistic and ontological annotation.
- A hybrid annotation model which combines linguistic and ontological annotations and SCHUG (Declerck, 2002), with the results of the morphological annotation developed by the team led by Dr Aquilino Sánchez, professor of the University of Murcia;

- Design and implementation of the platform architecture for hybrid annotation. The design is already finished and the following modules have been implemented:
  - Development of ontologies for representing linguistic knowledge to be used for standardizing annotations. In our project we have followed the EAGLES standard.
  - First prototype where FDG (Tapanainen & Järvinen, 1997) results are integrated.
  - The annotations are provided in the XML language though they can be converted into RDF(S) or OWL, by reusing the translation services available in WebODE.

The following tasks are planned to be carried out next year:

- Homogenization and standardization of annotations using the proposed platform.
- Integration of the tools.
- Empirical comparison and evaluation of the results obtained by the different annotation tools integrated into the platform after being homogenized and standardized.
- Integrate OntoTag into WebODE.

The integration of different NL analysis tools (or the results provided by different analysers) and the development of linguistic ontologies (which were not counted on in the project proposal) have delayed this sub-plan and the following one.

Subplan III OntoConsult. Natural language interface based on ontologies.

We have designed and built a new interface (henceforth "semantic portal"), which permits consulting the contents of web pages in any domain taking into account the information supplied by an existing ontology (previously built) . We have called this technology SeW.

- The SeW architecture is built on top of the WebODE platform.
- The modules forming the interface have been designed taking into account that the semantic portal contents should be synchronized with the ontologies used by the semantic portal; therefore, if the ontology is modified with the WebODE editor, the changes will appear on the portal.
- A first version of SeW has already been implemented.

The following tasks are still pending for next year:

- To add further SeW functionalities.
- To develop a module that permits any user to consult web pages in natural language with the same analysers as those of sub-plan II.

Subplan IV OntoAdvice. A system based on ontologies for browsing and collecting information from web pages annotated in the entertainment domain.

This sub-plan has been started in the second year. As the technology developed in the previous three sub-plans are needed to carry out this application, the only task performed along the year has been the development of anontology in the film domain using for the Methontology methodology for ontology construction, and the WebODe platform.

More than 30 paper documents have been consulted to develope the ontology, among these are (Fernández-Tabau, 1994),(Costa, 1995), (Romaguera i Rambió, 1999),(Ferrando, 2000) and (Sánchez-Noriega, 2002). Different web sites have also been consulted, being *Internet Movie Data Base* (IMDb) the most relevant (http://www.cs.umbc.edu/~skallu1/IMDb.pdf), which holds a film ontology developed by the University of Maryland; this ontology is available in RDFS

(http://139.91.183.30:9090/RDF/VRP/Examples/imdb.rdf). It is a worldwide reference for the cinema lovers: It is visited over 13million times a month and holds more that 18.000 films classified. This ontology has been used as starting point to build our ontology. Other websites have also been visited to collect information, as for example, ThoughtTreasure of IBM (http://www.signiform.com/tt/htm/navext.htm), and The Greatest Films page (http://www.filmsite.org/), where documents created by experts on American and Hollywood films can be consulted/browsed.

In general, most of the difficulties arisen are related to the selection of lexical, morphological and syntactic issues and to the integration of the results supplied by the analysers available. Unfortunately, these difficulties have delayed sub-plans II and III. We might also find difficulties when analysing the questions made in natural language by users in sub-plans III and IV. It may happen that when this project is finished the module be extremely limited in its functionalities.

3 Results obtained

3.1. Personnel

One following final project has been read:

- *Ontología de Cine para la Web Semántica*, by **Francisco Rey Alamillo**. This final project was directed by Dr Mariano Fernández López, in February 2003 at the Facultad de Informática of the Universidad Politécnica de Madrid.

And other two are in progress:

- **Jaime Cantais** is developing a linguistic DB that store the results from OntoTag. Such linguisitic information is used by the ontology learning module for the authomatic construction of ontologies.
- **Javier Arrizabalaga** is implementing some modules of the OntoTag architecture.

The following **doctoral theses** are still in progress:

- *ODESeW. Generación Automática de Portales de Conocimientos para Intranets y Extranets.* Doctoral Thesis by **Angel López Cima** and directed by Dr Asunción Gómez Pérez.
- *Modelo de aprendizaje semi-automático de ontologías basado en el análisis de lenguaje natural*, Doctoral Thesis by **David Manzano Macho**, codirected by Dr Guadalupe Aguado de Cea and Dr Asunción Gómez Pérez.
- *OntoTag: Modelo de anotación lingüístico-ontológica en el contexto de la Web Semántica.* Doctoral Thesis by **Antonio Pareja Lora** . This thesis is codirected by Dr Guadalupe Aguado de Cea and Dr Asunción Gómez Pérez.

3.2. Research papers published

The tasks carried out have yielded the following papers:

1. Aguado-de Cea, G., Álvarez de Mon-Rego, I., Gómez-Pérez, A., Pareja-Lora, A. & Plaza-Arteche, R. (2002a) *A Semantic Web Page Linguistic Annotation Model. Semantic Web Meets Language Resources.* Technical Report WS-02-16, pp. 20-29. ISBN 1-57735-169-x. **American Association for Artificial Intelligence.** AAAI Press. Menlo Park, California, E.E.U.U.

2. Aguado, G., Álvarez-de-Mon, I., Pareja-Lora, A. & Plaza-Arteche, R. (2002b) *OntoTag: A Semantic Web Page Linguistic Annotation Model.* Proceedings of the **ECAI 2002 Workshop on Semantic Authoring, Annotation and Knowledge Markup.** Lyon, Francia.
3. Aguado, G., Álvarez-de-Mon, I., Pareja-Lora, A. & Plaza-Arteche, R. (2002c) *RDF(S)/XML linguistic annotation of Semantic Web pages.* **Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002). COLING'2002.** Taipei, Taiwan.
4. Aguado-de Cea, G., Álvarez de Mon-Rego, I., Gómez-Pérez, A., Pareja-Lora, A. (2003b) *OntoTag: XML/RDF(S)/OWL Semantic Web Page Annotation in ContentWe*b. **Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2003) – Language Technology and the Semantic Web,** pp. 25-32. 10th Conference of the European Chapter of the Association for Computational Linguistics. EACL'03. Budapest, Hungría.
5. Aguado de Cea, G., Álvarez de Mon y Rego, I., Pareja Lora, A. (2003c) *Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag.* Inteligencia Artificial, **Revista Iberoamericana de Inteligencia Artificial**. No.18, pp. 37–49. ISSN 1137 - 3601.
6. Arpírez, J.; Corcho, A.; Fernández-López, M.; Gómez-Pérez, A. WebODE in a Nutshell. **AI Magazine.** Fall 2003. Paginas. 37-47.
7. Buitelaar, P., Bryant, B., Ide, N., Lin, J., Pareja-Lora, A., Wilcock, G. (2003) *The Roles of Natural Language and XML in the Semantic Web.* Language and Linguistics (en prensa)
8. Fernández-López, M., Gómez-Pérez, A (2002) *The integration of OntoClean in WebODE.* **EKAW2002 Workshop on Evaluation of Ontology Tools (EON2002).** Sigüenza. September 2002.
9. Corcho O, Fernández-López M, Gómez-Pérez A (2003) *Methodologies, tools and languages for building ontologies. Where is the meeting point.* **Data and Knowledge Engineering,** 46:41-64
10. Corcho O, Fernández-López M, Gómez-Pérez A, Vicente O (2002) *WebODE: an Integrated Workbench for Ontology Representation, Reasoning and Exchange.* In: Gómez-Pérez A, Benjamins VR (eds) **13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)**. Sigüenza, Spain. (Lecture Notes in Artificial Intelligence LNAI 2473) Springer-Verlag, Berlin, Germany, pp 138–153
11. Fernández-López M, Gómez-Pérez A (2002b) *The integration of OntoClean in WebODE.* In: Angele J, Sure Y (eds) **EKAW'02 Workshop on Evaluation of Ontology-based Tools** (EON2002), Sigüenza, Spain. CEUR Workshop Proceedings 62: 38–52. Amsterdam, The Netherlands (http://CEUR-WS.org/Vol-62
12. Corcho O, Gómez-Pérez A, *WebPicker: Knowledge Extraction from Web Resources. Lectures Notes in Informatics: Applications of Natural Language to Information Systems.* Ed Moreno, A.M.; van de Riet, R.P.**German Informatics Society.** Köllen Druck + Verlag GMBH, Bonn Junio 2001. Pag. 55-64.
13. Corcho O, Gómez-Pérez A, *Solving Integration Problems of e-commerce Standards and initiatives through ontological mappings.* **IJCAI'01 WS on Ontologies and Information Sharing.** CEUR Workshop Proceedings (CEUR-WS.org). (2001)

3.3. Technological Transference

In the near future we will try to contact with firms interested in these issues. Nevertheless, we have already commented on parts of the results with the following people and institutions:

- Richard Benjamins, from iSOCO.
- Antonio Valderrábanos, from **SEMA group**.

We have also contacted with the **ISO/TC 37/SC 4 committee,** "Language Resources Management", and particularly with the Working Group 2 (Representation Schemes) so that the linguistic ontologies and the annotation model generated in this project could be used for standardizing the definition of scheme representation and/or the morph-syntactical and syntactical annotation.

3.4. Patent: We have not patented any result yet.

3.5. Participation in EU project

Our project has permitted that some members of the research team participate in the following international projects on Ontologies – Ontology Learning, with techniques of natural language processing- and of the Semantic Web.

- **Esperonto Services**. Application Service Provision of Semantic Annotation, Aggregation, Indexing and Routing of Textual, Multimedia, and Multilingual WebContent
  IST-2001-34373 (FP5)
  Unión Europea (Action Line IST-2001-3.4.1)
- **OntoWeb:** Ontology-based information exchange for knowledge management and electronic commerce
  IST-2000-29243 (FP5)
  Unión Europea (Action Line IST-2000-3.4)

In addition, some members of the research team participated with two proposals of FP6 NoE (1st call): Knowledge Web y Language Web.

- The **Knowledge Web** (http://www.synapticworkgroup.com/us/sdk/index.html) proposal has already been approved and deals with ontologies and the semantic web. Dr Gómez Pérez is the Scientific Deputy Director of the NoE.
- The *Language Web* proposal has been rejected. This proposal lead by Paul Buitelard and Nicoletta Calzolari dealt with the application of natural language processing technology to the semantic web.

Dr Gómez Pérez has also participated in the creation of an Integrated Project, called **KEMIME**. This project combines the analysis and generation of natural language with ontologies. This proposal was ranged among the six first proposals presented.

3.5. Collaboration with national and foreign research teams.

Our project has given us the opportunity of collaborating with national and international working groups.

National Collaborations:

- We have worked with the team led by **Dr Aquilino Sánchez**, professor of Universidad de Murcia. With this group we have interchanged elements of the corpus built for the development of the project.
- With **Dr Teresa Cabré** and Instituto Universitario de Lingüística Aplicada (IULA), U. Pompeu Fabra, for reusing the tools developed by both our teams in new projects.
- With **Dr Ricardo Mairal**, professor of UNED, (Universidad Nacional de Educación a Distancia) to explore the semantic annotation models.
- **Universidad Politécnica de Cataluña** provided us a licence for using EuroWordNet in Spanish.
- With **Dr Manuel Lama Penin**, from the Universidad de Santiago de Compostela, to include new services in WebODE related to the semantic web services.

Internacional Participations:

- Asunción Gómez (lecturer), Guadalupe Aguado and Antonio Pareja (attendees) to the "***Final ISLE/EAGLES Workshop***", 2-3 December, 2002: The focus of this workshop was centred on the standardization of multilingual lexical resources and most specifically, on the collaborative work with other scientific communities (for ontology developing, Semantic Web, etc.)
- Presentation of the project and exhibition of the first results obtained in **SIG5** (*Special Interest Group on Language Technology in Ontology Development and Use*) of the European Thematic Net IST 2000-29243 *"OntoWeb: Ontology-based Information Exchange for Knowledge Management and Electronic Commerce"*. This group is now analyzing the application of the existing linguistic resources to the Semantic Web.
- Antonio Pareja Lora, member of the research team of this project has participated as a **committee member of the program *EACL 2003 Workshop on Language Technology and the Semantic Web – 3rd Workshop on NLP and XML (NLPXML-2003)***, sponsored/backed by SIG5 of *OntoWeb*.
- Collaboration with the ***University of Saarbrücken*** and ***DFKI*** in Germany. We are now integrating the results obtained with the SCHUG tool, developed by both institutions within the OntoTag platform.
- Participation with **Dr Nuria Bel** from the University of Barcelona to standardize the linguistic labels within the ISO/TC 37/SC 4 committee ,"Language Resources Management", particularly with the Working Group 2 (Representation Schemes).
- Collaboration with **Dr Valentina Tamma**, from the University of Liverpool, to use the techniques for analyzing the semantic distance between terms in the learning of ontologies.